DOCUMENT RESUME

ED 173 427 TM 009 541

AUTHOR Tollefson, Nona; Tracy, D. B.
TITLE Response Length and Quality in the Grading of Essay Tests.
PUB DATE Apr 79
NOTE 16p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 9-11, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Essays; *Essay Tests; *Evaluation Criteria; *Response Style (Tests); *Scoring; Secondary School Teachers; Senior High Schools; Social Studies; Student Teachers; *Test Reliability; Test Results; Test Validity
IDENTIFIERS *Test Length

ABSTRACT

The validity and reliability of essay scores were examined by comparing the mean scores assigned to good and poor quality essay responses of different lengths written by high school sophomores. In-service and pre-service social studies teachers graded essay responses to a test question requiring knowledge of the Constitutional provisions for police power and the relationship of police power to the 14th Amendment. A scoring key was used. Mean scores were analyzed using t-tests. Comparisons between responses of different qualities and different lengths indicated that the long response had a significantly higher mean score than short or moderate length responses; that good quality papers were graded significantly higher than poor quality responses; and that the mean for the poor quality long response was not significantly different from the mean responses for good quality short and moderate length responses. Results indicated that the length of the response--rather than knowledge of the subject--influenced scores, and that a scoring key alone was not sufficient to insure that quality of response was the main determinant of the grade assigned. (Author/MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED173427

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Response Length and Quality in the Grading of Essay Tests

Nona Tollefson and D. B. Tracy

University of Kansas

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Nona Tollefson
D.B. Tracy

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Presented to the National Council on Measurement in Education,

San Francisco, April, 1979.

TM009 541

Abstract

Response Length and Quality in the Grading of Essay Responses

The reliability and validity of essay scoring was investigated by comparing the mean scores assigned to good and poor quality essay responses of different lengths. In-service and pre-service social studies teachers graded a set of 10 responses to a social studies essay question. The long responses had a significantly higher means than responses of short or moderate length. Good quality responses were graded significantly higher than poor quality responses, and the mean for the poor quality long response was not significantly different from the mean for the good quality responses of short or moderate length.

Test Length and Quality in the Grading of Essay Responses

Since essay tests are relatively easy to construct and duplicate and represent a very direct and seemingly valid method of testing recall and integration of information, they are attractive to busy teachers. Although essay tests have the advantage of permitting each student to organize and present material in his/her own way, problems in assigning reliable and valid grades to essay responses have been documented in many tests. (Aiken, 1976; Ebel, 1972; Karmel & Karmel, 1978; Sax, 1972, Scannell & Tracy, 1975) If the grading process does not result in the reliable assignment of scores, the validity and consequent usefulness of essay tests is also questionable. To the greatest extent possible, scores assigned must reflect paper quality and be independent of irrelevant factors such as length, handwriting, and eccentricities of the grader.

Recent studies by Hales and Tolar (1975) and Tracy and Tollefson (1977) have shown that the order in which essay responses are read by the grader affects the grades assigned to moderate quality papers. The present study continued the investigation of factors which influence essay scoring. Essay responses of good and poor quality, but of different lengths were graded by in-service and pre-service teachers. The general hypothesis tested was that the quality rather than the length of the response determines the grade assigned.

Procedure

One method of studying the validity of grades assigned essay tests is to compare the grades assigned to the same paper by different competent

graders. (Scannell and Tracy, 1975, pg. 89). The present study compared the grades assigned by 88 competent graders to a set of essay responses which varied in length and quality.

The actual responses of 30 high school sophomores to the item: "Briefly explain and give examples of the term police power and the relation it has to the due process clause of the 14th amendment", were submitted to a panel of high school social studies teachers. These teachers graded the responses using a key which had been developed by the classroom teacher for assigning 0 to 10 points to each response. Means and standard deviations were computed for each paper. Papers were then classified into 2 groups as either poor or good quality papers using mean grades. Each group of papers was then subdivided by length of response into long, moderate, or short groups. This process resulted in six groups of papers.

Ten of the original 30 papers were selected with papers with the smallest variance in rating being retained within the six length-quality categories. Initially, the plan was to select 2 good and 2 poor quality papers from each response length group. However, the original set papers included only one long response which had been rated as poor, so only 2 long responses were included among the experimental papers. Thus the experimental set of papers included 4 short papers, 2 of which were of good quality, 4 moderate length papers, 2 were of good quality, and 2 long responses, 1 good and 1 poor quality. As the major concern of the study, was the effect of response length on grades, papers were arranged within experimental set so that a maximum number of different length, different quality comparisons could be made. Given this criterion, the experimental set of papers were presented to graders in the order shown in Table 1.

Insert Table 1 here

The experimental papers with the scoring key were presented to 88 in-service and pre-service social studies teachers with directions to award 0 to 10 points to each response.

Data Analysis and Results

The means and standard deviation for each paper and for the combined good and poor long, moderate, and short papers are reported in Table 2.

Insert Table 2 here

Repeated measures t-tests were used to test for differences between the mean scores for papers of different lengths and between the mean scores of papers of different quality. Each t-test was performed at the .0033 level of significance to control the experiment-wise Type I error rate at .05 or less.

Comparisons between the combined means of papers of different qualities are shown in Table 3. This analysis showed that the mean score for long

Insert Table 3 here

responses was significantly higher than those for short or moderate length responses and that the mean score for moderate length responses was significantly higher than the mean for short responses. Mean score for good quality responses was significantly higher than those for poor quality responses.

Since significant differences were found between the mean scores assigned groups of papers of different lengths and between the mean scores of papers of different quality, comparisons were performed between the means of papers of different lengths, but the same quality.

As shown in Table 4, when all papers were of good quality, long responses were rated significantly higher than short or moderate length responses. The finding was identical for poor quality responses. Comparisons between short and moderate length papers yielded a significantly higher mean for moderate length poor papers than for the short papers when both were of poor quality. The comparison between good quality papers of short versus moderate length was not significant.

Insert Table 4 here

Since longer responses were rated significantly higher than short or moderate length responses independent of the quality, comparisons among the poor quality long response and individual good quality short and moderate length responses were conducted. As shown in Table 5, *t*-tests were conducted between the mean long poor quality response (3) and the mean responses for each of the 4 good quality short and moderate length papers. These were papers 2 (moderate good), 4 (short good), 7 (short good), and 10 (moderate good).

Insert Table 5 here

All comparisons yielded non-significant results indicating that the mean ratings for the good quality papers were not significantly different from the mean rating assigned the long poor quality paper.

Discussion

In-service and pre-service teachers gave significantly higher grades to long essay responses than to short or moderate length essay responses regardless of the quality of the responses. The good quality long response had the highest mean rating and the smallest variance. The mean score of the poor quality long response did not differ significantly from the mean score for short or moderate length responses of good quality. These findings raise serious questions about the validity of the scoring procedures presently used by many teachers.

The sample of teachers participating in this study was not randomly selected from a defined population of sophomore social studies teacher trainees, but was probably a representative sample of this group. Assuming the findings for this group mirror the findings of other teacher groups, there is a need for improved scoring of essay test responses. For the group of student studied, the marks assigned to the responses indicated in large part how long the response was rather than the knowledge of the question asked.

The methodology used in this study provided that the teacher use a scoring key in marking responses, commonly recommended procedure for scoring essay responses. However, results of this study, indicate that a scoring key alone is not sufficient to insure that quality of response is the determinant of the grade assigned.

Summary

This study investigated the validity of essay scores by comparing the mean scores assigned to good and poor quality essay responses of different lengths. In-service and pre-service social studies teachers graded essay responses to a test question requiring knowledge of the Constitutional provisions for police power and the relationship of police power to the 14th Amendment. Mean scores were analyzed using t-tests. Comparisons between responses of different qualities and different lengths were conducted. Data analysis indicated that the long response had a significantly higher mean score than short or moderate length responses, that good quality papers were graded significantly higher than poor quality responses, and that the mean for the poor quality long response was not significantly different from the mean responses for good quality short and moderate length responses.

Reference Note

1. Tracy, D. B. and Tollefson, Nona. Effect of order of paper presentation on creative writing scores. Paper presented at the meeting of the Southwestern Psychological Association, Ft. Worth, Texas, April 1977.

References

Aiken, L. R. Jr. Psychological Testing and Assessment. Boston: Allyn and Bacon, Inc., 1976.

Ebel, R. L. Essentials of Educational Measurement. Englewood Cliffs, New Jersey: Prentice-Hall Inc., 1972.

Hales, L. W. and Tolar, E. The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. Journal of Educational Measurement, 1975, 12, 115-117.

Karmel, L. J. and Karmel, M. O. Measurement and Evaluation in the Schools, second edition. New York: MacMillan, 1978.

Sax, G. Principles of Educational Measurement and Evaluation. Belmont, California: Wadsworth, 1974.

Scannell, D. P. and Tracy, D. B. Testing and Measurement in the Classroom. Boston: Houghton Mifflin Co., 1975.

Table 1

Arrangement of Experimental Set of Paper by Length and Quality

Order	Length	Number of Words	Quality
1.	Short	47	Poor
2.	Moderate	104	Good
3.	Long	145	Poor
4.	Short	52	Good
5.	Moderate	105	Poor
6.	Long	144	Good
7.	Short	55	Good
8.	Moderate	100	Poor
9.	Short	35	Poor
10.	Moderate	93	Good

Table 2

Means and Standard Deviations for the Experimental Papers and the Combined Good and Poor Long, Moderate, and Short Papers

Individual Papers

Order	Length	Quality	Mean	Standard Deviation	n = 88
1.	Short	Poor	4.318	2.337	
2.	Moderate	Good	6.443	2.454	
3.	Long	Poor	6.409	2.043	
4.	Short	Good	5.943	2.511	
5.	Moderate	Poor	5.727	2.105	
6.	Long	Good	9.295	1.776	
7.	Short	Good	7.159	2.149	
8.	Moderate	Poor	5.830	2.075	
9.	Short	Poor	4.511	2.344	
10.	Moderate	Good	6.284	2.459	

Papers Combined for Length

Type of Response	Mean	Standard Deviation
Short	5.48	2.77
Moderate	6.07	2.49
Long	7.85	2.43

Papers Combined for Quality

Quality	Mean	Standard Deviation
Good	7.02	2.49
Poor	4.21	1.52

Papers Combined for Quality and Length

Quality	Length	Mean	Standard Deviation
Good	Short	6.55	3.84
Good	Moderate	6.36	4.07
Good	Long	9.30	3.15
Poor	Short	4.41	3.92
Poor	Moderate	5.78	2.69
Poor	Long	6.41	4.18

Table 3

Results of the Analysis of Mean Differences in Scores Assigned Groups of papers of different length and different quality

(Means are included in parentheses)

Comparison	t (df = 87)
Short (5.48) vs Long (7.85)	-12.38*
Short (5.48) vs Moderate (6.07)	- 3.56*
Moderate (6.07) vs Long (7.85)	-11.20*
Good (7.02) vs Poor (4.21)	19.00*

*p < .0033

Table 4

Results of t-tests for Mean Differences in Sets of Papers of same Quality but Different Lengths (Means are included in parentheses)

Quality	Comparison	t (df = 87)
Good	Short (6.55) vs Long (9.30)	-13.66*
Good	Short (6.55) vs Moderate (6.36)	.85
Good	Long (9.30) vs Moderate (6.36)	13.44*
Poor	Short (4.41) vs Long (6.41)	7.45*
Poor	Short (4.41) vs Moderate (5.78)	6.92*
Poor	Long (6.41) vs Moderate (5.78)	3.098*

*p < .0033

Table 5

Results of the t-tests for Comparisons between the Poor Quality Long Responses and Good Quality Short and Moderate Length Responses

Comparisons	t (df = 87)
Long poor (3) Short good (4)	-1.59
Long poor (3) Short good (7)	-2.65
Long poor (3) Moderate good (2)	- .12
Long poor (3) Moderate good (10)	.45